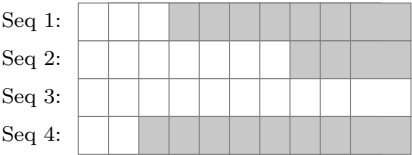


1. Naive Padding



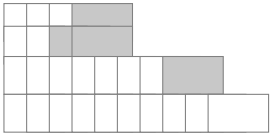
Memory Usage:
40 positions
60% waste

2. Length Bucketing

Bucket 1 (len ≤ 3):

Bucket 2 (len ≤ 7):

Bucket 3 (len ≤ 10):



Memory Usage:
24 positions
25% waste

3. Dynamic Batching

Batch 1:

Batch 2:

Batch 3:



Memory Usage:
23 positions
15% waste

Real Token

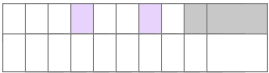
Padding

s Separator

4. Packed Sequences

Pack 1:

Pack 2:



Memory Usage:
20 positions
5% waste

5. Variable Length (Future)

Batch:



Memory Usage:
20 positions
0% waste

Performance Comparison

